Stability of neural networks and solitons of field theory

Sang-Hee Han and In Gyu Koh
Department of Physics, Korea Advanced Institute of Science and Technology (KAIST), Yusong-ku Daejon 305-701, South Korea
(Received 3 March 1999; revised manuscript received 7 September 1999)

The layers of a feed-forward neural network are interpreted as a cascade of field theories. The stability of the neural network is interpreted as the topological stability of kink solutions. An explicit example is shown for a three-class problem with their field theoretical Lagrangian equations. [S1063-651X(99)10112-0]

PACS number(s): 87.18.Sn, 84.35.+i, 05.45.Yv, 11.10.Ef

I. INTRODUCTION

Neural networks have been very successfully applied in the pattern recognition of handwritten characters, for example, in the postal service and of speech. A neural network is a system of layers in which the output of each layer, O_i, is given by a threshold function $g(x)$ [1]:

$$O_i = g\left(\sum_j \omega_{ij} x_j - \Theta_i \right), \tag{1}$$

where ω_{ij} represents the connection strength for an input node x_j of each layer and $g(x)$ is an activation function which shows a saturation nonlinearity, for instance,

$$g(x) = \tanh(\beta x).$$

For high β (~ 10), $g(x)$ can be regarded as a steplike function. After the weights have been updated by the well-known back-propagation algorithm [2], using suitable training data, the neural network is stable against fluctuations of the input data.

Variations in handwriting styles, which are absent in the training data, can be absorbed by the feed-forward process of the multilayer neural network. The capacity of abstracting from the training data and of generalizing to handle the difference between the training data and new data is one of the most prominent features of neural networks. The neural network system is stable against variations in input, such as writing styles and different individual characteristics, and is thus able to give correct classifications.

In this paper, we examine this stability from the viewpoint of field theory and its soliton solutions. Each layer of a feed-forward neural network corresponds to a mapping of the space of lower nodes onto the fields of upper nodes. The fields of the lower layer become the space upon which the new fields of the next layer are formulated. This concept of several layers has analogies in string theories where the fields defined over two-dimensional space-time become new space-time upon which the usual gauge and gravitational fields are formulated.

II. FIELD THEORETICAL APPROACH TO FEED-FORWARD NETWORKS

A feed-forward neural network consists of several layers where the input data of each layer are related to the output O_i as in Eq. (1). In our field theory model, the value of a node on a layer is denoted as $\phi(x)$, a function of the node values x of the preceding layer. ϕ is of dimension m, which is the number of nodes in the upper layer, while the x is of dimension n, which is the number of nodes in the lower layer.

We wish to search for an activation function in the field theory that saturates at $+$ $(-)$ 1 and favors large arguments in order to enhance the stability of the pattern. We start from a generic field Lagrangian [3–5]

$$\mathcal{L} = -\frac{1}{2} \nabla \phi \cdot \nabla \phi - V(\phi) \tag{2}$$

and interpret ∇ as a vector derivative with respect to lower nodes variables. The potential $V(\phi)$ is

$$V(\phi) = \sum_i \left[\left(\sum_j \omega_{ij}^2 \right) \left(\frac{1}{2} \phi_i^4 - \phi_i^2 \right) \right]. \tag{3}$$

The coefficients

$$-\frac{1}{2} \left(\sum_j \omega_{ij}^2 \right)$$

in the potential are the inverse square of the length scales of the model.

Since the lower nodes have the same signature, we have not included the kinetic terms in Eq. (2). The equations of motion are obtained by the Euler-Lagrange equation

$$\sum_j \frac{\partial}{\partial x_j} \frac{\partial \phi_i}{\partial x_j} = \left(\sum_j \omega_{ij}^2 \right) 2(\phi_i^3 - \phi_i). \tag{4}$$

The potential in Eq. (3) has a double minimum for one ϕ_i as shown in Fig. 1. The value of minima (3) is

$$-\frac{1}{2} \left(\sum_j \omega_{ij}^2 \right),$$

where the ground state of this potential is either $\phi_i = 1$ or -1. Another interesting solution is

$$\phi_i = \tanh(y_i) = \tanh\left(\sum_j \omega_{ij} x_j - \theta_i \right), \tag{5}$$

where $S_i : \Sigma \omega_{ij} x_j - \theta_i = 0$ is the equation for the separating plane in the feature space x_j of dimensions n. The kink so-

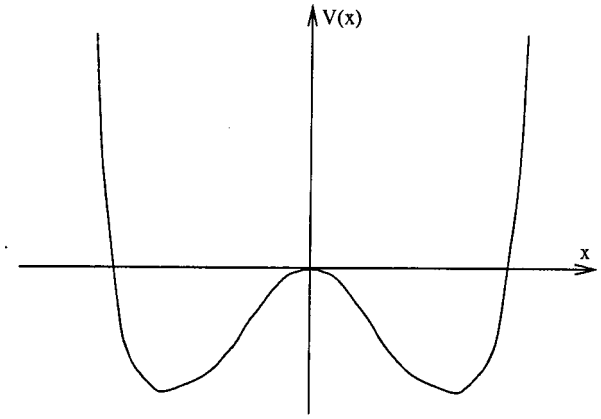

FIG. 1. The potential for the field ϕ_i which is the node value of upper layer in the neural network. The minimum of potential is $\phi_i=1$ or -1.

lution (5) is -1 on the left side of the separating boundary and $+1$ on the right side of the boundary. This solution is stable topologically, since the right-hand side $+1$ would require an infinite amount of energy to be transformed into a -1 state. The shape of the soliton solutions depends on the coefficients ω_{ij}, which are determined by the training data used in the back-propagation algorithm.

The minimum action principle for equations of motion requires the integral of Eq. (2) over x to be minimal. This minimum action may suggest that the sum of the square of errors over large training data with a different input feature x is required to be minimum for optimized neural networks.

It may be appropriate to interpret the field theoretical Lagrangian equations from the neural network viewpoint to understand the interesting relationship between field theory and the feed-forward neural system. For a multilayer network, every nth layer $\phi^{(n)}$ can be regarded as the input x for the $(n+1)$th layer and can be mapped onto the next layer $\phi^{(n+1)}$. The $\nabla\phi$ term in Eq. (2) is related to the change in the output layer upon small changes in the input feature x. The potential (3) is suggested for the output layer to have the desired classification in the neural network. In the region of feature space x where classifications are very stable, the first term in Eq. (2) and changes of Eq. (3) for small variations are small, but near the boundary of classification they are of significant magnitudes.

Thus stability in the classification of the pattern is achieved by the kink solution (5) where the input variation of the lower planes does not produce changes. Although the soliton solution of the Euler-Lagrange equation is not a unique activation function for general neural networks, choosing a soliton solution as an activation function provides the appropriate topology to the feature space x and, consequently, always gives a stable feature to the general classification problem that we want to apply.

III. EXAMPLE

To gain a better understanding of the relationship between field theory and neural networks, let us observe some general features from this explicit numerical work. According to the argument of the previous section, other functions which do not satisfy the variation principle for the Lagrangian equation would have nonminimal action. Nonoptimal activation functions, for example, stretched step functions, with a linear range have been observed to have a larger sum of square errors.

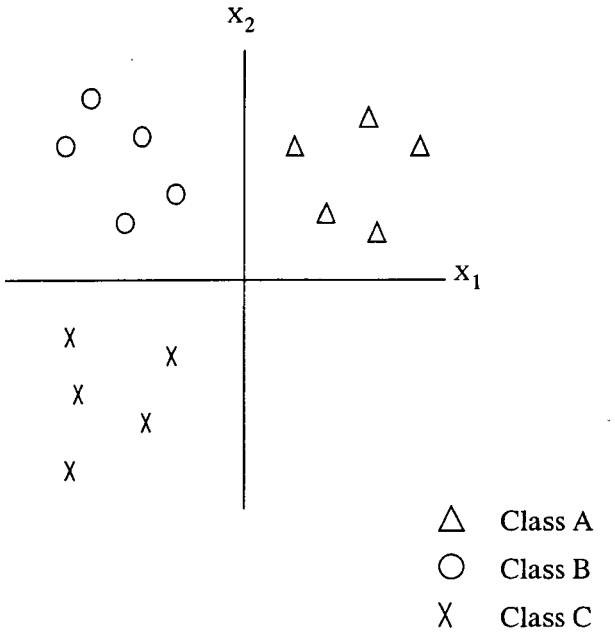

FIG. 2. Training data for the example three-class problem with two-dimensional input features.

We will demonstrate the field theoretical approach of neural networks with an example of the three-class problem with two-dimensional features. The back-propagation algorithm used training data in Fig. 2, and the resulting network in Fig. 3 has parameters

$$w^{(1)}=\begin{pmatrix}0.13 & 1.97\\ 3.22 & 0.03\end{pmatrix},\quad \theta^{(1)}=\begin{pmatrix}-0.26\\ -0.09\end{pmatrix},$$

$$w^{(2)}=\begin{pmatrix}0.51 & 2.35\\ 2.22 & -2.19\\ -2.41 & -0.59\end{pmatrix},\quad \theta^{(2)}=\begin{pmatrix}0.50\\ 2.18\\ 0.60\end{pmatrix}.$$

The weights $w^{(n)}$ in the nth layer connect the node $x_i^{(n)}$ of the nth layer with the node $x_j^{(n-1)}$ of the $(n-1)$th layer. The shifts $\theta^{(n)}$ in the nth layer are for the node i.

The nodes $x_1^{(1)}$ and $x_2^{(1)}$ in the first layer are the fields ϕ_1 and ϕ_2. The nodes $x_1^{(0)}$ and $x_2^{(0)}$ are regarded as x_1 and x_2 variables, respectively, for the fields ϕ_1 and ϕ_2. The field Lagrangian equation for ϕ_1 and ϕ_2 is, from Eqs. (2) and (3),

$$\mathcal{L}=-\frac{1}{2}(\nabla\phi_1\nabla\phi_1+\nabla\phi_2\nabla\phi_2)-3.90\left(\frac{1}{2}\phi_1^4-\phi_1^2\right)$$

$$-10.37\left(\frac{1}{2}\phi_2^4-\phi_2^2\right). \tag{6}$$

The two-dimensional ϕ_1,ϕ_2 have kink solutions

$$\phi_1=\tanh(0.13x_1+1.97x_2+0.26),$$

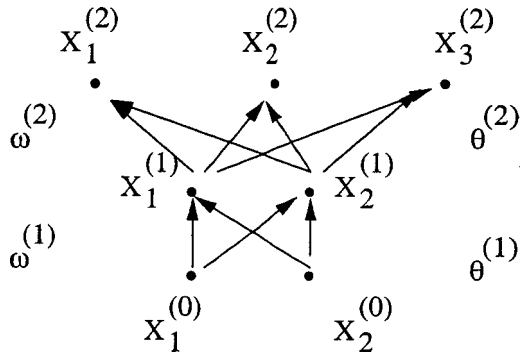

FIG. 3. The architecture of the neural network for the three-class problem with two-dimensional input features. The weights ω and the thresholds θ are trained by the back-propagation algorithm. The weights $\omega^{(2)}$ between the nodes $x^{(2)}$ and $x^{(1)}$ are defined in 3×2 matrices, and the $\theta^{(2)}$ values for the node $x^{(2)}$ are shown in 3×1 matrices. Similarly, the weights $\omega^{(1)}$ between the nodes $x^{(1)}$ and $x^{(0)}$ and $\theta^{(1)}$ are defined in 2×2 and 2×1 matrices.

$$\phi_2=\tanh(3.22x_1+0.03x_2+0.09),$$

in the variables of x_1 and x_2, and their characteristic separate planes are shown in Fig. 4. The physical meaning of separating planes becomes clear. The three classes of pattern are separated by the topology of the planes, and the classifying sample input data only depend on their sign. Near the separating plane, the kink solutions have values between -1 and 1. Far away from the separating plane boundary, they have vacuum degeneracy.

These ϕ_1,ϕ_2 become the variables x_1,x_2 for the second layer, and the three nodes $x_1^{(2)},x_2^{(2)},x_3^{(2)}$ in the second layer become the fields ϕ_1,ϕ_2,ϕ_3. The corresponding Lagrangian equation in the second layer is

$$\mathcal{L}=-\frac{1}{2}(\nabla\phi_1\nabla\phi_1+\nabla\phi_2\nabla\phi_2+\nabla\phi_3\nabla\phi_3)$$
$$-5.78\left(\frac{1}{2}\phi_1^4-\phi_1^2\right)-9.72\left(\frac{1}{2}\phi_2^4-\phi_2^2\right)$$
$$-6.16\left(\frac{1}{2}\phi_3^4-\phi_3^2\right). \qquad (7)$$

The kink solutions are

$$\phi_1=\tanh(0.51x_1+2.35x_2-0.50),$$

$$\phi_2=\tanh(2.22x_1-2.19x_2-2.18),$$

$$\phi_3=\tanh(-2.41x_1-0.59x_2-0.60),$$

and their behaviors are shown in Fig. 5.

The test data A_1, A_2, B_1, B_2, C_1, and C_2 for the three classes *A*, *B*, and *C* and their convergence are shown in Fig. 4. By the kink solutions of Eq. (6), the fluctuations in the input features belonging to the same class are removed. The kink solution absorbs the difference in test data A_1 and A_2. For example, A_1 and A_2 with the input features (1.0,3.2) and (3.0,3.0) are mapped onto (1.00,1.00) in Fig. 4. Also by the kink solutions of Eq. (7), both A_1 and A_2 are mapped onto (0.98, -0.97, -1.00) close to the ideal value (1.00, -1.00, -1.00). The input fluctuations are removed by the topological stability and are mapped onto the stable representative points.

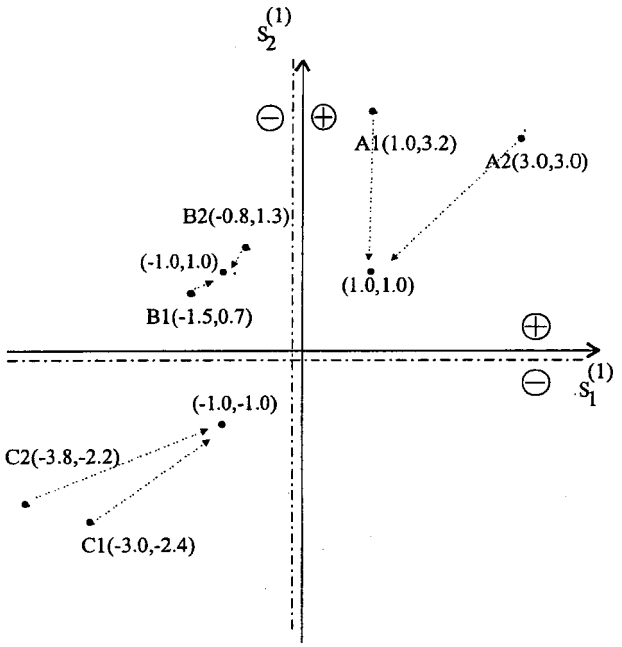

FIG. 4. The separating planes in the input feature spaces $x_1^{(0)}$ and $x_2^{(0)}$ are shown. The kink solutions have 1 (-1) in the right (left) side of the separating plane. Also, the sample data A_1, A_2, B_1, B_2, C_1, and C_2 for the next layer are shown.

IV. CONCLUSION

We interpreted the generalization power of a neural network as the topological stability of the classical solution of field theory. The feed-forward feature of neural networks is modeled as a cascade of field maps. The categorizing method to use a mapping to field space has been very successful both in explaining the intrinsic discrete feature of neural network layers and, in practical manner, in absorbing fluctuations of

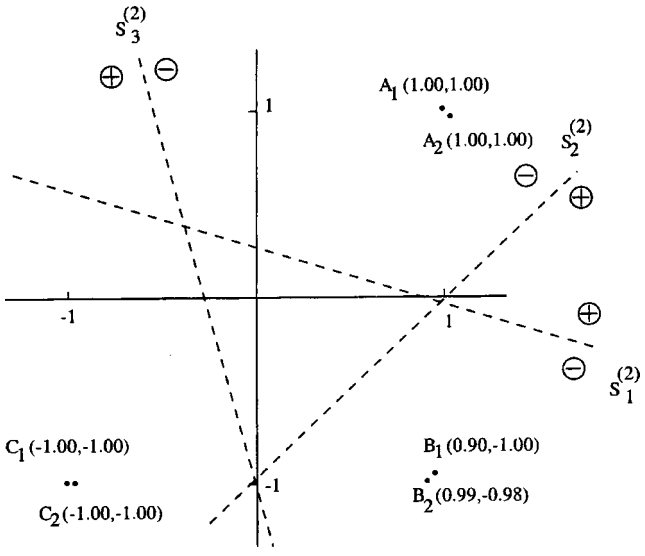

FIG. 5. The separating planes in the node values $x_1^{(1)}$ and $x_2^{(1)}$.

input data features and in giving stability in the pattern classification to the network system. The stability of the system against fluctuations is achieved either by a differential equation with negative eigenvalues or by topological stability. The neural network has been studied from the viewpoint of differential equations. However, the discrete nature of a layer of a neural network is significantly different from the behavior of a differential equation. On the other hand, the topological stability of neural networks enables us to understand how input variations are eliminated. The cascade behavior of neural network repeats the mapping from the space to the field. In each successive layer, the field of the lower layer becomes the space variable and the next field theory based on the space variable is constructed. A similar construction has been observed in string field theory [6]. When the input space is two dimensional, it is called field theory. The fields in the two dimensions become those space-time variables, and one constructs anomaly-free supergravity models based on those space-time variables. The specific choice of a potential and its solution need to be studied further.

[1] W. S. McCulloch and W. Pitts, Bull. Math. Biophys. **5**, 115 (1946).

[2] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control* (Blaisdell, New York, 1969).

[3] S. Coleman, *Aspects of Symmetry* (Cambridge University Press, Cambridge, England, 1985), pp. 185–192.

[4] J. D. Bjorken and Sidney D. Drell, *Relativistic Quantum Fields* (McGraw-Hill, New York, 1965).

[5] C. Itzykson and J. B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York, 1980).

[6] M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory* (Cambridge University Press, Cambridge, England, 1987), Vol. 1.